МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
Национальный исследовательский ядерный университет «МИФИ»

Институт интеллектуальных кибернетических систем

КАФЕДРА КИБЕРНЕТИКИ

БДЗ

по курсу "Математическая статистика"

студента группы Б20-504

Яковлева Андрея

Вариант № 12

Оценка: _____

Подпись: _____

2021 г.

1. Описательные статистики

1.1. Выборочные характеристики

Анализируемый признак 1 – Appraisal price3 (D3)

Анализируемый признак 2 – Nitric oxides concentration (D9)

Анализируемый признак 3 – Weighted distances to five city employment centres (D10)

а) Привести формулы расчёта выборочных характеристик

Выборочная хар-ка	Формула расчета
Объём выборки	n
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Выборочная дисперсия	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$
Выборочное среднеквадратическое отклонение	$S = \sqrt{S^2}$
Выборочный коэффициент асимметрии	$A_s = \frac{m_3}{S^3}$, где $m_3 = \frac{1}{n} * \sum_{i=1}^{n}(x_i - \bar{x})^3$
Выборочный эксцесс	$E_k = \frac{m_4}{S^4}$, где $m_4 = \frac{1}{n} * \sum_{i=1}^{n}(x_i - \bar{x})^4$

б) Рассчитать выборочные характеристики

Выборочная хар-ка	Признак 1	Признак 2	Признак 3
Среднее	1260.444	6.230	4.779
Выборочная дисперсия	150972.870	0.196	2.571
Выборочное среднеквадратическое отклонение	388.552	0.443	1.603
Выборочный коэффициент асимметрии	1.349	1.229	0.479
Выборочный эксцесс	1.374	2.767	-0.142

1.2. Группировка и гистограммы частот

Анализируемый признак – Appraisal price3 (D3)

Объём выборки – 117

а) Выбрать число групп

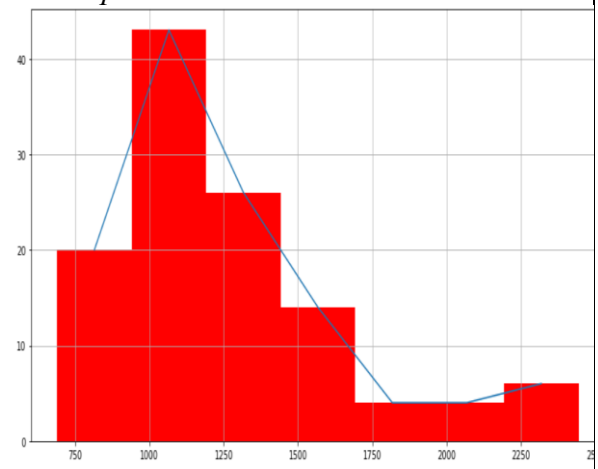
Число групп	Обоснование выбора числа групп	Ширина интервалов
7	Формула Скотта: $h = 3.5Sn^{-\frac{1}{3}}$ и $k = \frac{x_n - x_1}{h}$	277

б) Построить таблицу частот

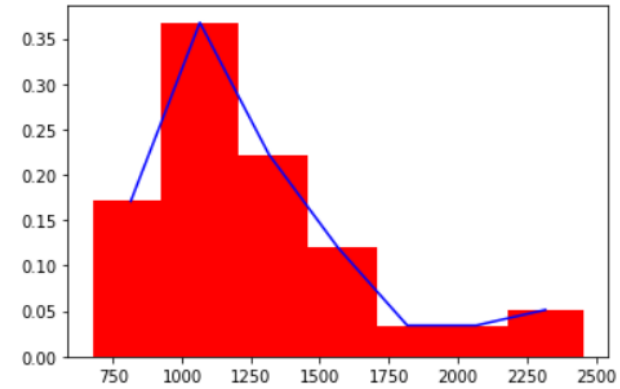
Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Накопл. частота	Относит. накопл. частота
1	690	967	24	0.205	24	0.205
2	967	1244	50	0.427	74	0.632
3	1244	1521	22	0.188	96	0.821
4	1521	1798	9	0.077	105	0.897
5	1798	2075	4	0.034	109	0.932
6	2075	2352	5	0.043	114	0.974
7	2352	2629	3	0.026	117	1

в) Построить гистограммы частот и полигоны частот

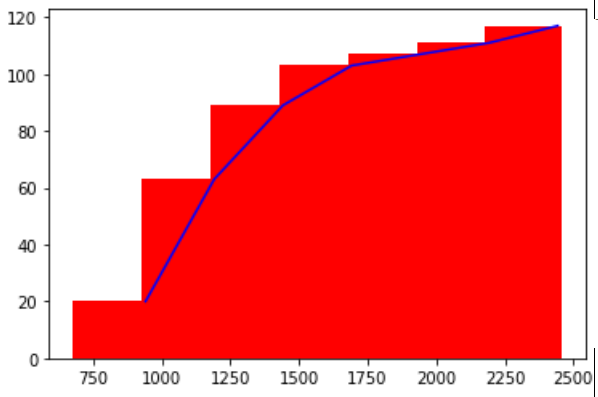
Гистограмма и полигон частот



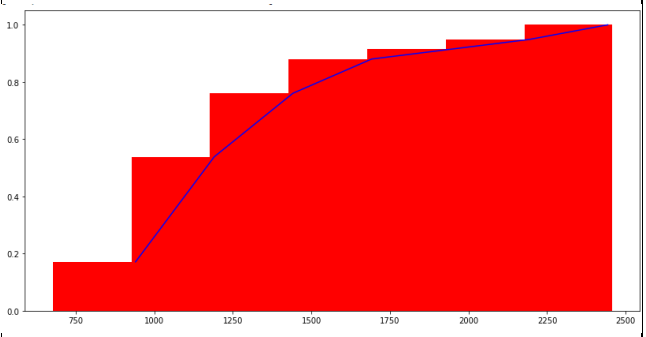
Гистограмма и полигон относительных частот



Гистограмма и полигон накопленных частот

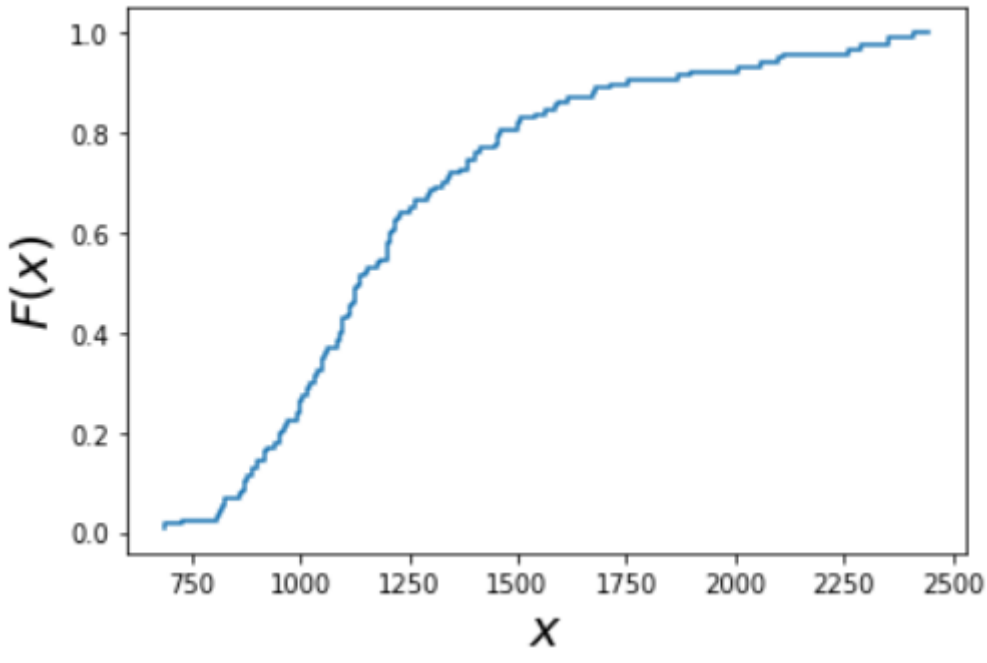


Гистограмма и полигон накопленных относительных частот



г) Построить график эмпирической функции распределения

Эмпирическая функция распределения



2. Интервальные оценки

2.1. Доверительные интервалы для мат. ожидания

Анализируемый признак – Appraisal price3 (D3)

Объём выборки – 117

Оцениваемый параметр – математическое ожидание

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\bar{X} - \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$
Верхняя граница	$\bar{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	1166.77	1189.6	1201.14
Верхняя граница	1354.11	1331.29	1319.75

2.2. Доверительные интервалы для дисперсии

Анализируемый признак – Appraisal price3 (D3)

Объём выборки – 117

Оцениваемый параметр – дисперсия

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}$
Верхняя граница	$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	109218.02	117556.83	122156.99
Верхняя граница	215632.25	197121.34	188456.11

2.3. Доверительные интервалы для разности мат. ожиданий

Анализируемый признак 1 – Appraisal price2 (D2)

Анализируемый признак 2 – Appraisal price3 (D3)

Объёмы выборок – 117

Оцениваемый параметр – m1 - m2

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$(\bar{X}_1 - \bar{X}_2) - t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Верхняя граница	$(\bar{X}_1 - \bar{X}_2) + t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	-224.467	-192.969	-176.952
Верхняя граница	36.501	5.003	-11.014

2.4. Доверительные интервалы для отношения дисперсий

Анализируемый признак 1 – Appraisal price2 (D2)

Анализируемый признак 2 – Appraisal price3 (D3)

Объёмы выборок – 117

Оцениваемый параметр – $\frac{\sigma_1^2}{\sigma_2^2}$

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{S_1^2}{S_2^2} f_{\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)$
Верхняя граница	$\frac{S_1^2}{S_2^2} f_{1-\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	0.6	0.68	0.72
Верхняя граница	1.58	1.4	1.32

3. Проверка статистических гипотез о математических ожиданиях и дисперсиях

3.1. Проверка статистических гипотез о математических ожиданиях

Анализируемый признак – Appraisal price3 (D3)

Объём выборки – 117

Статистическая гипотеза – $\begin{matrix} H_0 : m = m_0 \\ H' : m \neq m_0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$\frac{\bar{X} - m_0}{S/\sqrt{n}}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$T(n-1)$
Формулы расчета критических точек	$\pm t_{1-\frac{\alpha}{2}}(n-1)$
Формула расчета *p-value*	$\min\left(F_z(z), 1 - F_z(z)\right) * 2$

б) Выбрать произвольные значения m_0 и проверить статистические гипотезы

m_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
1100	0.1	4.467	$1.85 * 10^{-5}$	*H*0 отвергается	$m \neq 1100$
1400	0.1	-3.885	0.00017	*H*0 отвергается	$m \neq 1400$
1250	0.1	0.291	0.7717	*H*0 принимается	$m = 1250$

3.2. Проверка статистических гипотез о дисперсиях

Анализируемый признак – Appraisal price3 (D3)

Объём выборки – 117

Статистическая гипотеза – $\begin{matrix} H_0 : \sigma = \sigma_0 \\ H' : \sigma \neq \sigma_0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$\frac{(n-1)S^2}{\sigma_0^2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(n-1)$
Формулы расчета критических точек	$\chi^2_{\alpha/2}(n-1)$ $\chi^2_{1-\alpha/2}(n-1)$
Формула расчета *p-value*	$\min\left(F_z(z), 1 - F_z(z)\right) * 2$

б) Выбрать произвольные значения σ_0 и проверить статистические гипотезы

σ_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
320	0.1	301.444	*0.017*	*H*0 отвергается	$\sigma \neq 320$
340	0.1	434.079	*0.035*	*H*0 отвергается	$\sigma \neq 340$
390	0.1	114.156	*0.938*	*H*0 принимается	$\sigma = 390$

3.3. Проверка статистических гипотез о равенстве математических ожиданий

Анализируемый признак 1 – Appraisal price2 (D2)

Анализируемый признак 2 – Appraisal price3 (D3)

Объёмы выборок – 117

Статистическая гипотеза – $\begin{matrix} H_0 : m_1 = m_2 \\ H' : m_1 \neq m_2 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$

Закон распределения статистики критерия при условии истинности основной гипотезы	$N(0,1)$
Формулы расчета критических точек	$\pm N_{1-\frac{\alpha}{2}}(0,1)$
Формула расчета *p-value*	$\min\left(F_z(z), 1 - F_z(z)\right) * 2$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	1.863	0.064	$H0$ принимается	$m_1 = m_2$
0.05			$H0$ принимается	$m_1 = m_2$
0.1			$H0$ отвергается	$m_1 \neq m_2$

3.4. Проверка статистических гипотез о равенстве дисперсий

Анализируемый признак 1 – Appraisal price2 (D2)

Анализируемый признак 2 – Appraisal price3 (D3)

Объёмы выборок – 117

Статистическая гипотеза – $\begin{matrix} H_0 : \sigma_1 = \sigma_2 \\ H' : \sigma_1 \neq \sigma_2 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$\frac{S_1^2}{S_2^2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(n_1 - 1, n_2 - 1)$
Формулы расчета критических точек	$F_{\frac{\alpha}{2}}(n_1, n_2); F_{\left(1-\frac{\alpha}{2}\right)}(n_1, n_2)$
Формула расчета *p-value*	$\min\left(F_z(z), 1 - F_z(z)\right) * 2$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	1.028	0.883	$H0$ принимается	$\sigma_1 = \sigma_2$
0.05			$H0$ принимается	$\sigma_1 = \sigma_2$
0.1			$H0$ принимается	$\sigma_1 = \sigma_2$

4. Критерии согласия

Анализируемый признак – Appraisal price3 (D3)

Объём выборки – 117

4.1. Критерий хи-квадрат

Теоретическое распределение – нормальное.

Статистическая гипотеза – $H_0 : F(x)$ □ N

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$	k-число интервалов ni – число элементов в i-м интервале
Закон распределения статистики критерия при условии истинности основной гипотезы	χ^2 (k − r − 1)	pi - вероятности попадания в i-й интервал при условии истинности $H0$
Формула расчета критической точки	$\chi^2_{1-a}(k - r - 1)$	n – объем выборки
Формула расчета *p-value*	$1 - \chi^2(z, k - r - 1)$	r – число неизвестных параметров распределения

б) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
7	$h = 3.5Sn^{-\frac{1}{3}}$ и $k = \frac{x_n - x_1}{h}$	277

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Вероятность попадания в интервал при условии истинности основной гипотезы
1	690	967	24	0.205	0.15
2	967	1244	50	0.427	0.26
3	1244	1521	22	0.188	0.27
4	1521	1798	9	0.077	0.17
5	1798	2075	4	0.034	0.065
6	2075	2352	5	0.043	0.015
7	2352	2629	3	0.026	0.0021

г) Построить гистограмму относительных частот и функцию плотности теоретического распределения на одном графике

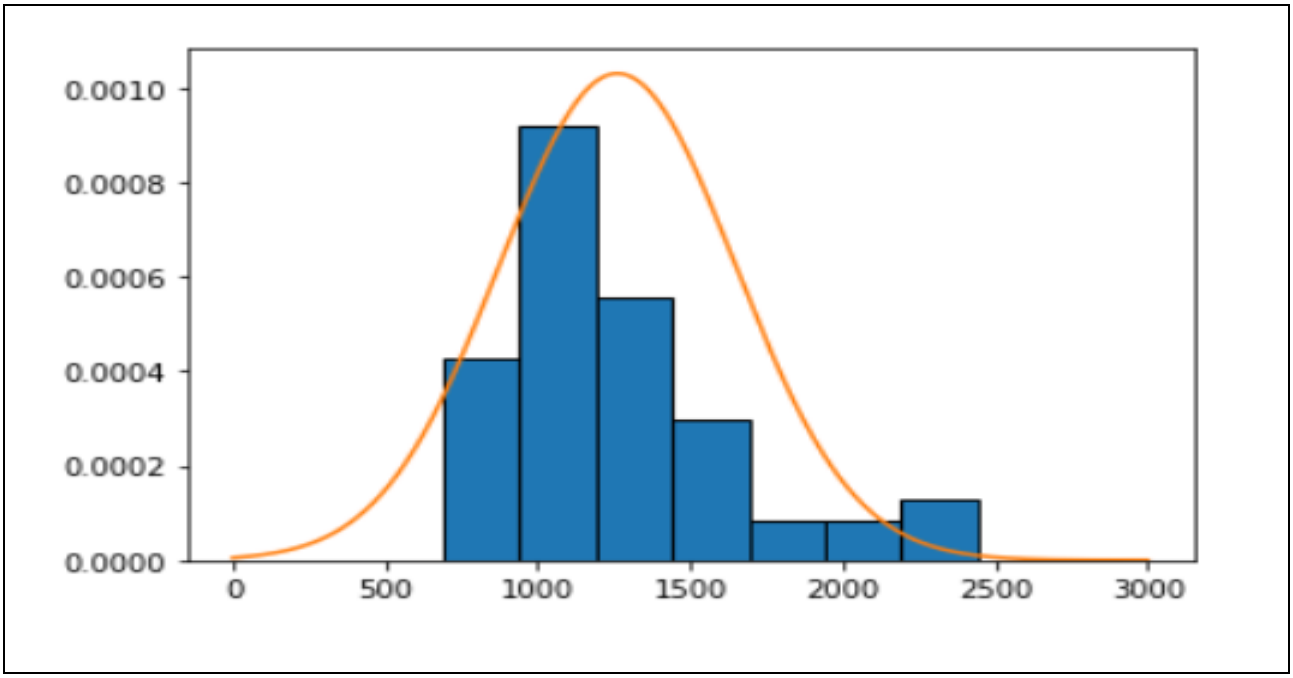


График относительных частот уменьшен в 400 раз.

д) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	53.132	$7.99 * 10^{-11}$	H_0 отвергается	F(D3) ∉ N
0.05	53.132	$7.99 * 10^{-11}$	H_0 отвергается	F(D3) ∉ N
0.1	53.132	$7.99 * 10^{-11}$	H_0 отвергается	F(D3) ∉ N

4.2. Проверка гипотезы о нормальности на основе коэффициента асимметрии и эксцесса (критерий Харке-Бера)

Статистическая гипотеза – $H_0 : F(x) \square N$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$n(\frac{S^2}{6} + \frac{K^2}{24})$	n – объем выборки
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(2)$	$S = \frac{\mu_3}{\sigma^3}$ – коэффициент асимметрии
Формула расчета критической точки	$\chi^2_{1-a}(2)$	$K = \frac{\mu_4}{\sigma^4} - 3$ - коэффициент эксцесса
Формула расчета *p-value*	$1 - \chi^2(z, 2)$	μ_i- i-й центральный момент σ – стандартное отклонение

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	46.639	$7.45 * 10^{-11}$	H_0 отвергается	F(D3) ∉ N
0.05			H_0 отвергается	F(D3) ∉ N
0.1			H_0 отвергается	F(D3) ∉ N

Вывод (в терминах предметной области)

В результате проведённого в п.4 статистического анализа обнаружено, что оценочная стоимость домов (D3) не является нормально распределенной величиной.

5. Проверка однородности выборок

Анализируемый признак 1 – Appraisal price3 (D3)

Анализируемый признак 2 – Appraisal price2 (D2)

Объёмы выборок – 117

5.1 Критерий знаков

Статистическая гипотеза – $H_0 : F_1(x) = F_2(x)$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$2\sqrt{n}$ (H – 0.5)	n – объем выборок
Закон распределения статистики критерия при условии истинности основной гипотезы	$N(0,1)$	H = K/n – частота успеха
Формула расчета критической точки	$\pm \mathrm{N}_{1-a/2}$ (0,1)	K – число знаков «+» в последовательности знаков разностей
Формула расчета *p-value*	$\min\left(F_z(z), 1 - F_z(z)\right) * 2$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	8.23	$2.22 * 10^{-16}$	H_0 отвергается	$F_1(x) \neq F_2\ (x)$
0.05			H_0 отвергается	$F_1(x) \neq F_2\ (x)$
0.1			H_0 отвергается	$F_1(x) \neq F_2\ (x)$

5.2. Критерий хи-квадрат

Статистическая гипотеза – $H_0 : F_1(x) = F_2(x)$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$n_X n_Y \sum_{i=1}^{k} \frac{1}{m_i^X + m_i^Y} (\frac{m_i^X}{n_X} - \frac{m_i^Y}{n_Y})^2$	n1 – число элементов в 1-ой выборке
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2\ (k\ -\ 1)$	n2 – число элементов во 2-ой выборке
Формула расчета критической точки	$\chi^2_{1-a}(k-1)$	${m^X}_i$ – частота первой выбороки в i – й группе
Формула расчета *p-value*	$1 - \chi^2\ (z, k-1)$	${m^Y}_i$ – частота второй выбороки в i – й группе

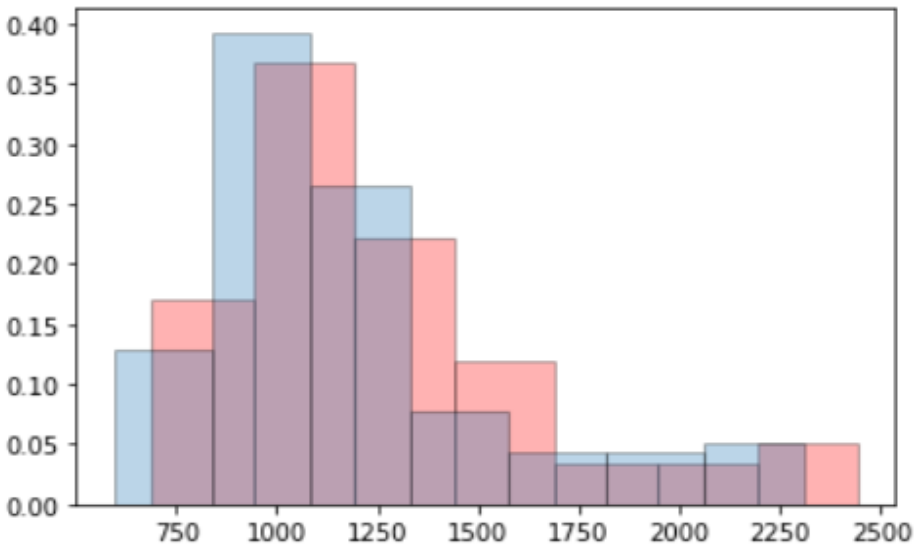
б) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
7	$k\ =\ [1\ + log_2 n]$	

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота признака 1	Частота признака 2	Относит. частота признака 1	Относит. частота признака 2
1	594	940.4	20	36	0.17	0.307
2	940.4	1190.9	43	41	0.37	0.35
3	1190.9	1441.29	26	21	0.22	0.18
4	1441.29	1691.71	14	6	0.12	0.05
5	1691.71	1942.14	4	4	0.03	0.034
6	1942.14	2192.57	4	4	0.03	0.034
7	2192.57	2443	6	5	0.05	0.043

г) Построить гистограммы относительных частот на одном графике



д) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	2.56	0.633	H_0 принимается	$F_1(x) = F_2(x)$
0.05			H_0 принимается	$F_1(x) = F_2(x)$
0.1			H_0 принимается	$F_1(x) = F_2(x)$

Вывод (в терминах предметной области)

В результате проведённого в п.5 статистического анализа не удалось определить, имеют ли выборки оценочная стоимость домов D2 и оценочная стоимость домов D3 одинаковые распределения. Критерий хи-квадрат принял гипотезу однородности. Критерий знаков отклонил ее. Следовательно, в одной из проверок была допущена ошибка, либо первого рода в критерии знаков, либо ошибка второго рода в критерии хи-квадрат.

6. Таблицы сопряжённости

Факторный признак *x* – Corner location (D8)

Результативный признак *y* – Nitric oxides concentration (D12)

Объёмы выборок –117

Статистическая гипотеза – $H_0 : F_Y(y|X = x_1) = \cdots = F_Y(y|X = x_k) = F_Y(y)$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}$	K – кол-во вариантов факторного признака
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi 2((k-1)(l-1))$	l– кол-во вариантов результативного признака
Формула расчета критической точки	$\chi^2_{1-a}((k-1)(l-1))$	n_{ij}– частота в эмпирической т.с.
Формула расчета *p-value*	$1 - \chi 2(z, (k-1)(l-1))$	m_{ij}– частота в теоретической т.с.

б) Построить эмпирическую таблицу сопряжённости

x \ y	N	Y	Σ
N	67	23	90
Y	12	15	27
Σ	79	38	117

в) Построить теоретическую таблицу сопряжённости

x \ y	N	Y	Σ
N	60.77	29.23	90
Y	18.23	8.77	27
Σ	79	38	117

г) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	8.52	0.004	H_0 отвергается	Между D8 и D12 есть статистическая связь
0.05			H_0 отвергается	Между D8 и D12 есть статистическая связь
0.1			H_0 отвергается	Между D8 и D12 есть статистическая связь

Вывод (в терминах предметной области)

В результате проведённого в п.6 статистического анализа обнаружено, что расположением угла (D3) и концентрацией оксидов азота (D12) есть статистическая связь.

7. Дисперсионный анализ

Факторный признак *x* – Location in sector of city (D7)

Результативный признак *y* – Appraisal price3 ($hundreds) (D3)

Число вариантов факторного признака – 3

Объёмы выборок – 117

Статистическая гипотеза – D7 влияет на D3

а) Рассчитать групповые выборочные характеристики

№ п/п	Вариант факторного признака	Объём выборки	Групповые средние	Групповые дисперсии
1	north	41	1462.610	179213.116
2	south	45	1125.356	88750.629
3	other	31	1189.161	113448.651

б) Привести формулы расчёта показателей вариации, используемых в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$D_b^* = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})^2$	K - 1	$\frac{n}{K-1}D_b^*$
Остаточные признаки	$D_\omega^* = \frac{1}{n}\sum_{k=1}^{K} n_k\widetilde{\sigma_k^2}$	n - K	$\frac{n}{n-K}D_\omega^*$
Все признаки	$D_X^* = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_i^k - \bar{x})^2$	n - 1	$\frac{n}{n-1}D_X^*$

в) Рассчитать показатели вариации, используемые в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	22687	2	1327214
Остаточные признаки	126995	114	129203
Все признаки	149682	116	150972

г) Проверить правило сложения дисперсий

Показатель	$D_{межгр}$	$D_{внутригр}$	$D_{общ}$	$D_{межгр} + D_{внутригр}$
Значение	22687	126995	149682	149682

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Эмпирический коэффициент детерминации	$\eta^2 = \frac{D_b^*}{D_x^*}$	0.15
Эмпирическое корреляционное отношение	$\eta = \sqrt{\frac{D_b^*}{D_x^*}}$	0.39

е) Охарактеризовать тип связи между факторным и результативным признаками

Между выборками D7 и D3 имеется умеренная связь.

ж) Указать формулы расчёта показателей, используемых при проверке статистической гипотезы дисперсионного анализа

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$\frac{D_b^*/(K-1)}{D_\omega^*/(n-K)}$	K – кол-во вар-в факторного признака
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(K-1, n-K)$	n-объем выборки
Формула расчета критической точки	$F_{1-\alpha}(K-1, n-K)$	D_b^* - межгрупповая дисперсия
Формула расчета *p-value*	$1 - F(z, K-1, n-K)$	D_ω^* - внутригрупповая дисперсия

з) Проверить статистическую гипотезу дисперсионного анализа

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	10.18	8.5e-05	H_0 отвергается	D7 влияет на D3
0.05			H_0 отвергается	D7 влияет на D3
0.1			H_0 отвергается	D7 влияет на D3

Вывод (в терминах предметной области)

В результате проведённого в п.7 статистического анализа обнаружено, что расположение дома относительно города (D7) оказывает влияние на оценочную стоимость дома (D3).

8. Корреляционный анализ

8.1. Расчёт парных коэффициентов корреляции

Анализируемый признак 1 – Appraisal price2 (D2)

Анализируемый признак 2 – Square feet of living space (D4)

Объёмы выборок – 117

а) Рассчитать точечные оценки коэффициентов корреляции

	Формула расчета	Значение
Линейный коэффициент корреляции	$\frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sigma_X{}^* \sigma_y{}^*}$	0.831
Ранговый коэффициент корреляции по Спирмену	$\frac{\sum_{i=1}^n (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^n (r_i - \bar{r})(s_i - \bar{s})}}$	0.853
Ранговый коэффициент корреляции по Кендаллу	$\frac{4Q}{n(n-1)} - 1$	0.665

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

Граница доверительного интервала	Формула расчета
Нижняя граница	$\rho_{XY}{}^* + \frac{\rho_{XY}{}^*(1-(\rho_{XY}{}^*)^2)}{n} - u_{1-\alpha/2}\frac{1-(\rho_{XY}{}^*)^2}{\sqrt{n}}$
Верхняя граница	$\rho_{XY}{}^* + \frac{\rho_{XY}{}^*(1-(\rho_{XY}{}^*)^2)}{n} + u_{1-\frac{\alpha}{2}}\frac{1-(\rho_{XY}{}^*)^2}{\sqrt{n}}$

в) Рассчитать доверительные интервалы для линейного коэффициента корреляции

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	0.74	0.77	0.78
Верхняя граница	0.89	0.88	0.87

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициентов корреляции

Статистическая гипотеза	Формула расчета статистики критерия	Закон распределения статистики критерия при условии истинности основной гипотезы

$H_0: \rho = 0$ $H': \rho \neq 0$	$\frac{\rho_{XY}^*}{\sqrt{1-(\rho_{XY}^*)^2}}\sqrt{n-2}$	$T(n-2)$
$H_0: r^{(сп)} = 0$ $H': r^{(сп)} \neq 0$	$\frac{\tilde{\rho}_{XY}^{(sp)}}{\sqrt{1-\left(\tilde{\rho}_{XY}^{(sp)}\right)^2}}\sqrt{n-2}$	$T(n-2)$
$H_0: r^{(кен)} = 0$ $H': r^{(кен)} \neq 0$	$\tilde{\tau}_{XY}\sqrt{\frac{9n(n+1)}{2(2n+5)}}$	$N(0, 1)$

д) Проверить значимость коэффициентов корреляции

Статистическая гипотеза	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
$H_0: \rho = 0$ $H': \rho \neq 0$	0.1	28.75	*0*	H_0 отвергается	$p \neq 0$
$H_0: r^{(сп)} = 0$ $H': r^{(сп)} \neq 0$	0.1	33.68	*0*	H_0 отвергается	$r^{(сп)} \neq 0$
$H_0: r^{(кен)} = 0$ $H': r^{(кен)} \neq 0$	0.1	10.72	*0*	H_0 отвергается	$r^{(кен)} \neq 0$

8.2. Расчёт множественных коэффициентов корреляции

Анализируемый признак 1 – Appraisal price2 (D2)

Анализируемый признак 2 –Square feet of living space (D4)

Анализируемый признак 3 – – D6 Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)

Объёмы выборок – 117

а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу

Признак / Признак	D2	D4	D6
D2	1	0.67	0.33
D4	0.67	1	0.304
D6	0.33	0.304	1

б) Рассчитать матрицу значений p-value для ранговых коэффициентов корреляции по Кендаллу (статистическая гипотеза $H_0: r^{(кен)} = 0,\ H': r^{(кен)} \neq 0$ *)*

Признак / Признак	D2	D4	D6
D2	–	0	1.44e-07
D4	0	–	9.25e-07

D6	1.44e-07	9.25e-07	–

в) Рассчитать точечную оценку коэффициента конкордации

	Формула расчета	Значение
Коэффициент конкордации	$\frac{12}{k^2(n^3-n)}\sum_{i=1}^{n}\left(\sum_{j=i}^{k}R_{ij}-\frac{k(n+1)}{2}\right)^2$ R_{ij} – – ранг i элемента в X_j выборке	12.895

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициента конкордации

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$n(k-1)W$	W – коэффициент конкардации
Закон распределения статистики критерия при условии истинности основной гипотезы	χ 2 $(n-1)$	n – размер выборок
Формула расчета критической точки	$\chi^2_{1-\alpha}(n-1)$	k – число выборок
Формула расчета *p-value*	$1-\chi^2(z,n-1)$	

д) Проверить значимость коэффициента конкордации

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	3017.43	*0.0*	H_0 отвергается	D2,D4,D6 - связаны
0.05			H_0 отвергается	D2,D4,D6 - связаны
0.1			H_0 отвергается	D2,D4,D6 - связаны

Вывод (в терминах предметной области)

В результате проведённого в п.8 статистического анализа обнаружено, что оценочная стоимость дома (D2), площадь дома (D4) и наличие признаков (D6) статистически связаны, в том числе и попарно.

9. Регрессионный анализ

9.1 Простейшая линейная регрессионная модель

Факторный признак *x* – Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access) (D6)

Результативный признак *y* – Appraisal price1 ($hundreds) (D1)

Уравнение регрессии – $f(x) = \beta_0 + \beta_1 x$

9.1.1. Точечные оценки линейной регрессионной модели

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
β_0	$\widetilde{\beta_0} = \bar{y} - \rho^*_{XY}\frac{\sigma^*_Y}{\sigma^*_X}\bar{x}$	661
β_1	$\widetilde{\beta_1} = \rho^*_{XY}\frac{\sigma^*_Y}{\sigma^*_X}$	114

б) Записать точечную оценку уравнения регрессии

$f(x) = 114x + 661$

в) Привести формулы расчёта показателей вариации, используемых в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$D^*_{Y\|X} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i, \beta_0, \dots \beta_{k-1}) - \bar{y})^2$	k - 1	$\frac{n}{k-1}D^*_{Y\|X}$
Остаточные признаки	$D^*_{resY} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, \beta_0, \dots \beta_{k-1}))^2$	n – k	$\frac{n}{n-k}D^*_{resY}$
Все признаки	$D^*_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$	n - 1	$\frac{n}{n-1}D^*_Y$

г) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка

Факторный признак	25345	1	2965347
Остаточные признаки	118150	115	120205
Все признаки	143495	116	144732

д) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	25345	118150	143495	143495

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$R_{Y\|X}{}^2 = \frac{D_{Y\|X}}{D_Y}$	0.176
Корреляционное отношение	$R_{Y\|X} = \sqrt{\frac{D_{Y\|X}}{D_Y}}$	0.42

ж) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

Связь, определяемая простейшей линейной регрессией – слабая.

9.1.2. Интервальные оценки линейной регрессионной модели

а) Привести формулы расчёта доверительных интервалов для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	Формула расчета
β_0	Нижняя граница	$\widetilde{\beta_0} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 D^*_X}}$
	Верхняя граница	$\widetilde{\beta_0} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 D^*_X}}$
β_1	Нижняя граница	$\widetilde{\beta_1} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\frac{1}{nD^*_X}}$

	Верхняя граница	$\widetilde{\beta_1} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\frac{1}{nD^*_X}}$

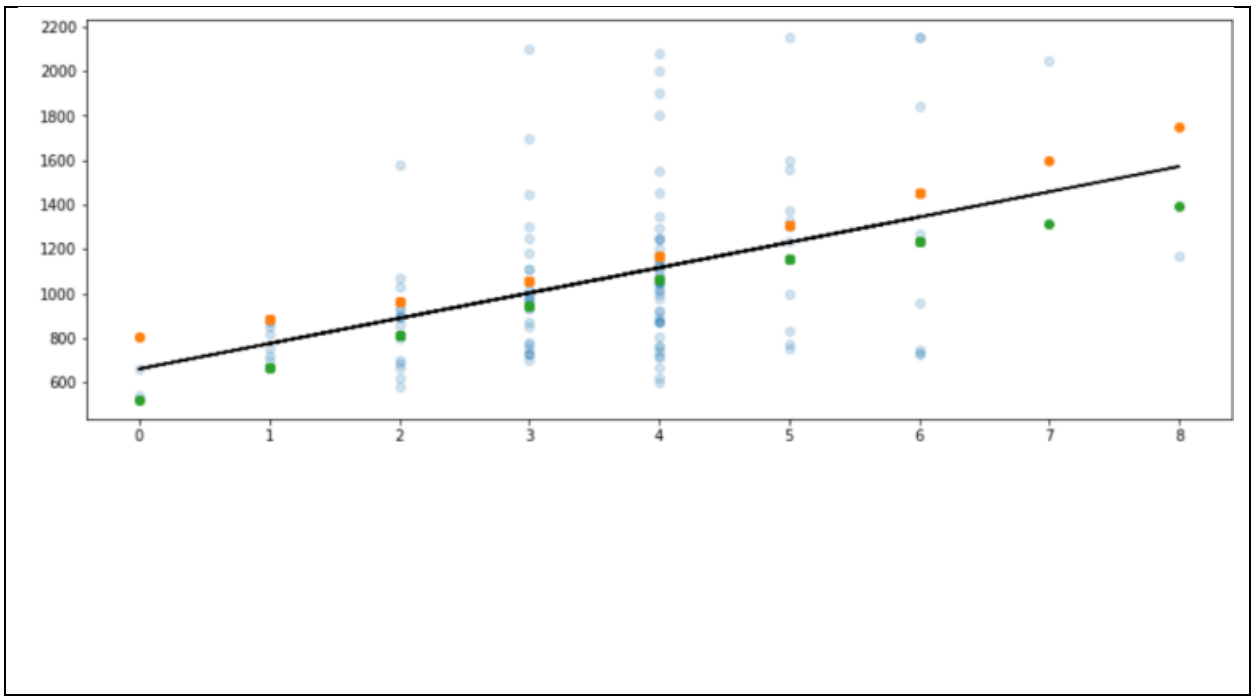
б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
β_0	Нижняя граница	435	490	518
	Верхняя граница	887	831	803
β_1	Нижняя граница	106	109	110
	Верхняя граница	121	118	117

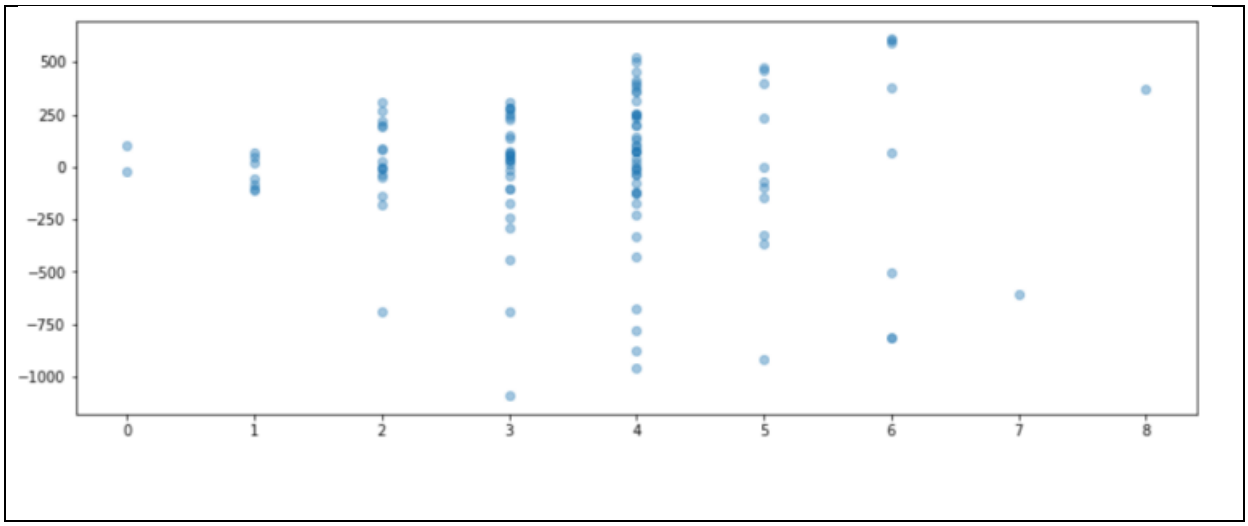
в) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$\widetilde{f(x)} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{n^2 D^*_X}}$
Верхняя граница $f_{high}(x)$	$\widetilde{f(x)} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{n^2 D^*_X}}$

г) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии $f_{low}(x)$ и $f_{high}(x)$ на уровне значимости $\alpha = 0.1$



д) Построить график остатков $\varepsilon(x) = y - f(x)$



9.1.3. Проверка значимости линейной регрессионной модели

Статистическая гипотеза – $H_0 : \beta_1 = 0$

$H' : \beta_1 \neq 0$

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

	Выражение	Пояснение использованных обозначений

Формула расчета статистики критерия	$\frac{R^{*}_{Y\|X}{}^{2}}{(1-R^{*}_{Y\|X}{}^{2})/(n-2)}$	$R^{*}_{Y\|X}{}^{2}$ - коэффициент детерминации
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(1, n-2)$	n – размер выборки
Формула расчета критической точки	$F_{1-\alpha}(1, n-2)$	
Формула расчета *p-value*	$1 - F(z, 1, n-2)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	24.7	2.4e-06	H_0 отвергается	Регрессионная модель значима
0.05			H_0 отвергается	Регрессионная модель значима
0.1			H_0 отвергается	Регрессионная модель значима

9.2 Линейная регрессионная модель общего вида

Факторный признак *x* – Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access) (D6)

Результативный признак *y* – Appraisal price1 ($hundreds) (D1)

Уравнение регрессии – квадратичное по *x*: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

9.2.1. Точечные оценки линейной регрессионной модели

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
β_0	$F = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \dots & \\ 1 & x_n & x_n^2 \end{pmatrix}$	640
β_1	$\tilde{\beta} = (F^T F)^{-1} F^T y$	128
β_2		-1.92

б) Записать точечную оценку уравнения регрессии

$f(x) = -1.92x^2 + 128x + 640$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	25381	2	1484787
Остаточные признаки	118114	114	121222
Все признаки	143495	116	144732

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	25381	118114	143495	143495

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$R_{Y\|X}{}^2 = \frac{D_{Y\|X}}{D_Y}$	0.176
Корреляционное отношение	$R_{Y\|X} = \sqrt{\frac{D_{Y\|X}}{D_Y}}$	0.42

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

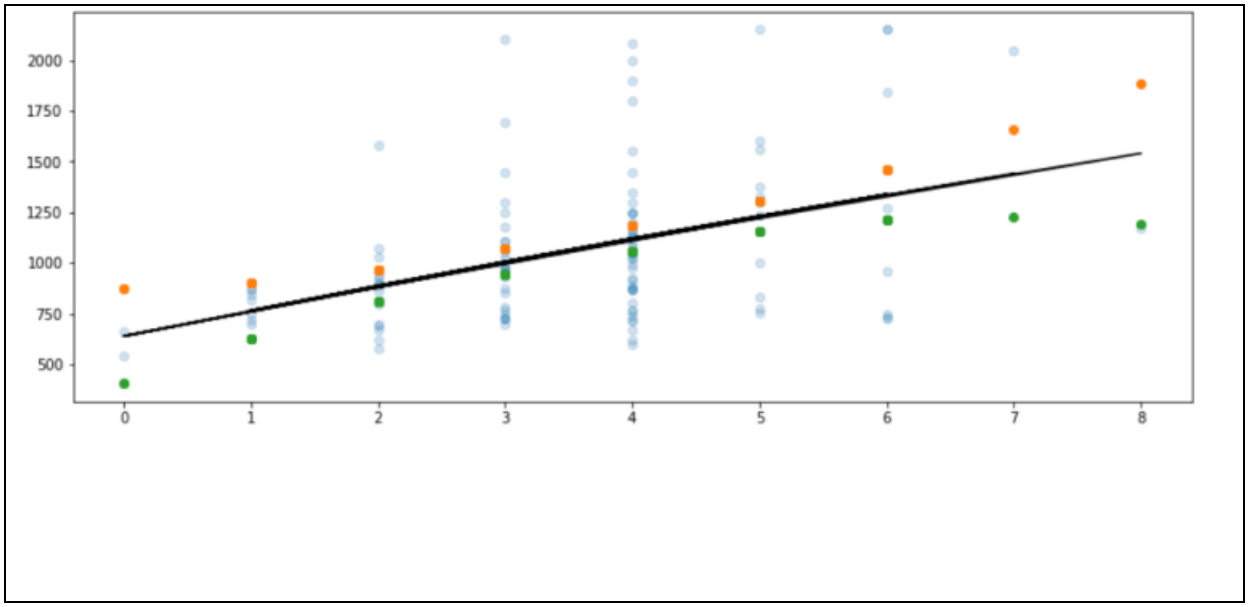
Связь, определяемая регрессией – слабая.

9.2.2. Интервальные оценки линейной регрессионной модели

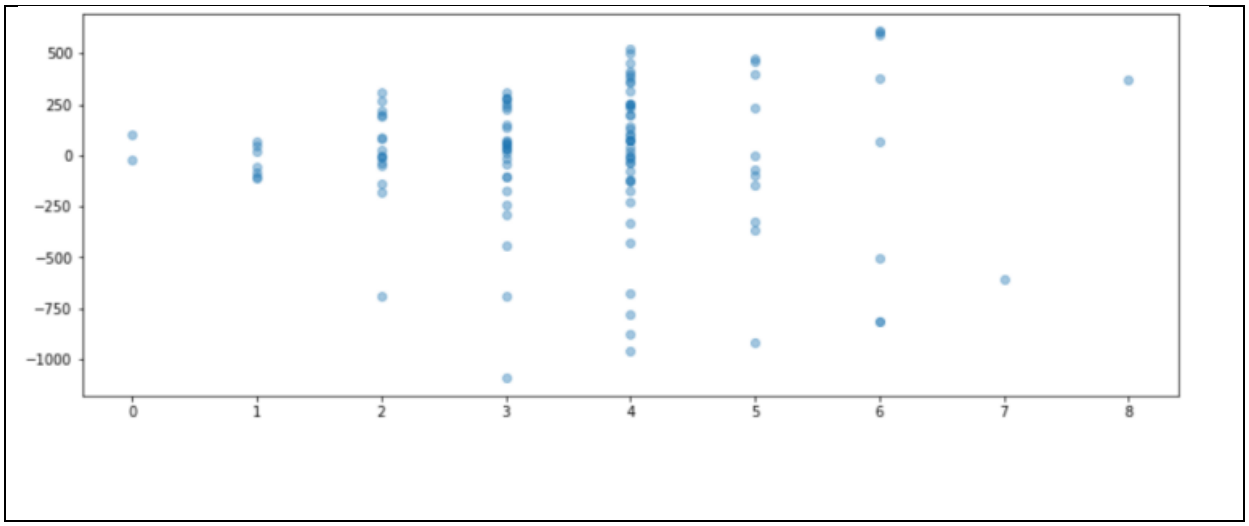
а) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$\widetilde{f(x)} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\varphi^T(x)(F^TF)^{-1}\varphi(x)}$
Верхняя граница $f_{high}(x)$	$\widetilde{f(x)} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{resY}}\sqrt{\varphi^T(x)(F^TF)^{-1}\varphi(x)}$

б) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии $f_{low}(x)$ и $f_{high}(x)$ на уровне значимости $\alpha = 0.1$



в) Построить график остатков $\varepsilon(x) = y - f(x)$



9.2.3. Проверка значимости линейной регрессионной модели

Статистическая гипотеза – $\begin{matrix} H_0 : \beta_1 = \beta_2 = 0 \\ H' : не\ H_0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$\frac{R^{*2}_{Y\|X}/(k-1)}{(1-R^{*2}_{Y\|X})/(n-k)}$	$R^{*2}_{Y\|X}$ – коэффициент детерминации
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(k-1, n-2)$	n – размер выборки
Формула расчета критической точки	$F_{1-\alpha}(k-1, n-2)$	
Формула расчета *p-value*	$1 - F(z, k-1, n-2)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	12.25	1.5e-05	H_0 отвергается	Регрессионная модель значима
0.05			H_0 отвергается	Регрессионная модель значима
0.1			H_0 отвергается	Регрессионная модель значима

9.3 Множественная линейная регрессионная модель

Факторный признак 1 x_1 – Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access) (D6)

Факторный признак 2 x_2 – Nitric oxides concentration (D9)

Результативный признак y – Appraisal price1 ($hundreds) (D1)

Уравнение регрессии – $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение

β_0	$F = \begin{pmatrix} 1 & x_1^1 & x_1^2 \\ 1 & x_2^1 & x_2^2 \\ & \dots & \\ 1 & x_n^1 & x_n^2 \end{pmatrix}$	-222
β_1	$\tilde{\beta} = (F^T F)^{-1} F^T y$	112
β_2		143

б) Записать точечную оценку уравнения регрессии

$f(x) = 112x_1 + 143x_2 - 222$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	29719	2	1738569
Остаточные признаки	113776	114	116770
Все признаки	143495	116	144732

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	29719	113776	143495	143495

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Множественный коэффициент детерминации	$R_{Y\|X_1,X_2}{}^2 = 1 - \frac{D_{res}}{D_Y}$	0.21
Множественное корреляционное отношение	$R^*_{Y\|X_1,X_2} = \sqrt{\frac{\left(\rho^*_{YX_1}\right)^2 + \left(\rho^*_{YX_2}\right)^2 - 2\rho^*_{YX_1}\rho^*_{YX_2}\rho^*_{X_1X_2}}{1 - \left(\rho^*_{X_1X_2}\right)^2}}$	0.45

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

Связь, определяемая регрессией – умеренная

9.4. Выводы

а) Сводная таблица показателей вариации для различных регрессионных моделей

Источник вариации	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Факторный признак	25345	25551	29716
Остаточные признаки	118150	117944	113779
Все признаки	143495	143495	143495

б) Сводная таблица свойств различных регрессионных моделей

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	17.6%	17.8%	21%
Значимость	Да	Да	Да
Адекватность	Нет	Нет	Нет
Степень тесноты связи	Слабая	Слабая	Умеренная

Вывод (в терминах предметной области)

В результате проведённого в п.9 статистического анализа обнаружено, что ни одна из предложенных регрессионных моделей адекватно не отражает реальную зависимость оценочной стоимости (D1) от наличия функций (D6) и от концентрации NO2 (D9).